A Simple Proof for the Optimality of Randomized Posterior Matching

Ofer Shayevitz and Meir Feder

Abstract

Posterior matching (PM) is a sequential horizon-free feedback communication scheme introduced by the authors, who also provided a rather involved optimality proof showing it achieves capacity for a large class of memoryless channels. Naghshvar *et al* considered a non-sequential variation of PM with a fixed number of messages and a random decision-time, and gave a simpler proof establishing its optimality via a novel Shannon-Jensen divergence argument. Another simpler optimality proof was given by Li and El Gamal, who considered a fixed-rate fixed block-length variation of PM with an additional randomization. Both these works also provided error exponent bounds. However, their simpler achievability proofs apply only to discrete memoryless channels, and are restricted to a non-sequential setup with a fixed number of messages. In this paper, we provide a short and transparent proof for the optimality of the fully sequential horizon-free PM scheme over general memoryless channels. Borrowing the key randomization idea of Li and El Gamal, our proof is based on analyzing the random walk behavior of the shrinking posterior intervals induced by a reversed iterated function system (RIFS) decoder.

I. Introduction

Posterior Matching (PM) is a simple and general feedback communication scheme introduced by the authors, who also showed it achieves capacity for a large class of memoryless channels, including discrete alphabets, continuous alphabets, and mixtures thereof [1]–[3]. One appealing feature of the PM scheme is that it is horizon-free and sequential, in the sense that the transmitter may send an infinite sequence of bits, and the receiver can decide to stop at every instant n; the receiver is then able decode roughly nC bits from the prefix of this sequence with vanishing error probability, where C is the capacity of the channel. Alternatively, the receiver is also able to decode the bits on the fly as soon as they become reliable enough. As argued in [1], PM can easily be converted to the more traditional settings where the number of messages and/or the horizon are fixed.

While heuristic arguments for the optimality of PM are simple and appealing (see [1], and going back to the special case of the Horstein scheme [4], [5]), the original optimality proof in [1] is quite involved and nontransparent. Coleman [6] studied the PM scheme from a novel stochastic control and Lyopanov exponent perspective, and provided a conceptually cleaner approach for its analysis. Naghshvar *et al* [7] considered a non-sequential variation of PM restricted to discrete memoryless channels (DMCs), where the number of messages is fixed but the decision time (horizon) is random. Introducing a novel Shannon-Jensen divergence, they provided a simpler proof showing that their scheme achieves the capacity of any DMC. Li and El Gamal [8] considered the same setting but with a fixed horizon. They described a randomized variation of PM and provided a simpler proof showing it achieves the capacity of any DMC. A key ingredient in their scheme was a random shift applied to the message point after each PM iteration, which circumvented some of the analysis obstacles. Both [7] and [8] also provide error exponent results.

In this paper, we adopt the random shift idea of Li and El Gamal, and consider a randomized version of the fully sequential horizon-free PM scheme. We provide a short and transparent optimality proof, showing that this scheme achieves the capacity for a very large class of memoryless channels, including all DMC and also many continuous alphabet and mixed alphabet channels. Our proof is based on analyzing the random-walk behavior of a reversed iterated function system (RIFS) decoder introduced in [1]. Unlike the deterministic PM scheme in [1], the combination of RIFS decoding and the random shift operation facilitates a much cleaner analysis and avoids the problem of fixed points that was a major obstacle in the original proof.

The authors are with the Department of EE–Systems, Tel Aviv University, Tel Aviv, Israel {ofersha@eng.tau.ac.il, meir@eng.tau.ac.il}.
The work of O. Shayevitz was supported by the Israel Science Foundation, grant no. 1367/14.

II. Preliminaries

A. Definitions and Basic Lemmas

Recall that a real-valued stochastic process T_n is called a submartingale if $\mathbb{E}(T_{n+1} \mid T^n) \geq T_n$ for any n. The following result is well known.

Lemma 1 (Martingale Convergence Theorem [9])**.** *Let T_n be a submartingale. If $\sup_n \mathbb{E}|T_n| < \infty$ then T_n convergence a.s. to some r.v. T and $\mathbb{E}|T| < \infty$.*

Let $g : [0,1] \mapsto \mathbb{R}$ be a Lebesgue measurable function. With some abuse of notations, we naturally extend g to operate on subsets of its domain in an element-wise fashion, namely $g(A) \triangleq \cup_{x \in A}\{g(x)\}$ for any set $A \subseteq [0,1]$. We write $|A|$ for the Lebesgue measure of the set A, whenever the former exists. Define the λ-smoothed derivative of g to be

$$D_\lambda[g(x)] \triangleq \frac{1}{\lambda} \left| g\left(\left[x - \tfrac{\lambda}{2}, x + \tfrac{\lambda}{2} \right] \bmod 1 \right) \right|,$$

where $t \bmod 1 \triangleq t - \lfloor t \rfloor$ is the modulo 1 operation.[1] Let

$$D[g(x)] \triangleq \limsup_{\lambda \to 0} D_\lambda[g(x)].$$

The following lemma is easily verified.

Lemma 2. *If $g(x)$ is differentiable at $x_0 \in (0,1)$ with a derivative $g'(x_0)$, then $D[g(x_0)] = |g'(x_0)|$. Furthermore, if g is absolutely continuous on $[0,1]$, then*

$$D_\lambda[g(x)] = \mathbb{E}\left| g'\left((x + Q_\lambda) \bmod 1 \right) \right|,$$

where $Q_\lambda \sim$ Unif$\left(\left[-\frac{\lambda}{2}, \frac{\lambda}{2}\right]\right)$.

Now, further define

$$\overline{D}[g(x)] \triangleq \sup_{\lambda \in (0,1)} D_\lambda[g(x)].$$

When g is absolutely continuous and monotonic (which will be our case of interest), then $\overline{D}[g(x)]$ is the maximal stretching of any symmetric interval (modulo 1) around x by g. The following lemma is a consequence of the Hardy-Littlewood maximal inequality [10], and states that $\overline{D}[g(x)]$ is unlikely to be too large, provided that g is well behaved. The proof is relegated to the appendix.

Lemma 3. *Let $g : [0,1] \mapsto \mathbb{R}$ be absolutely continuous on $[0,1]$, and $X \sim$ Unif$([0,1])$. Then for any $a > 0$,*

$$\Pr\left(\overline{D}[g(X)] > a \right) \leq 9a^{-1}\mathbb{E}\left| g'(X) \right|.$$

Remark 1. Note that if g is Lipschitz (which corresponds in the sequel to the case of discrete alphabet channels), then a stronger asymptotic statement trivially holds: $\Pr\left(\overline{D}[g(X)] > a \right) = 0$ for all a large enough.

Let $(X,Y) \sim P_{XY}$ be jointly distributed real-valued random variables. Let F_X be the c.d.f. of X, and F_X^{-1} be its functional inverse, generally defined by

$$F_X^{-1}(v) \triangleq \inf\{x : F_X(x) > v\}.$$

It is easy to verify (see e.g. [1]) that we can always define an auxiliary r.v. $\Theta \sim$ Unif$([0,1])$ such that $X = F_X^{-1}(\Theta)$. This induces a joint distribution $P_{\Theta XY}$. Let $F_{\Theta|Y}(\theta \mid y)$ denote the conditional c.d.f. of Θ given Y, also known as the *PM kernel* [1]. We will also be interested in the *inverse PM kernel* $F_{\Theta|Y}^{-1}(v \mid y)$, which is the functional inverse of the PM kernel w.r.t. θ [1].

[1]One may equivalently identify $[0,1)$ with the circle $\mathbb{R}/\mathbb{Z}$, in lieu of the modulo notation. The cyclic definition of the smoothed derivative takes care of what happens near the edges of the unit interval, and is essential for our purposes later due to the random shift. The definition (and associated results in this section) work with minor adaptations for any other interval domains (with the proper modulo) or when the domain is $\mathbb{R}$ (without the modulo).

In the remainder of the paper, we restrict our attention to the following family $\mathfrak{F}$ of all distributions P_{XY} admitting the following two properties:

(P1) $F_{\Theta|Y}(\theta \mid y)$ (resp. $F^{-1}_{\Theta|Y}(v \mid y)$) is absolutely continuous and strictly monotone in $\theta \in [0,1]$ (resp. $v \in [0,1]$) for P_Y-a.a. y.

(P2) There exists some $\delta > 0$ such that

$$\lim_{\lambda \to 0} \mathbb{E}|\log D_\lambda[F^{-1}_{\Theta|Y}(V \mid Y)]|^{2+\delta} < \infty,$$

where $Y \sim P_Y$ and $V \sim \text{Unif}([0,1])$ are independent, and the λ-smoothed derivative is taken w.r.t. v.

Remark 2. The family $\mathfrak{F}$ is quite rich and includes all discrete distributions, as well as many continuous and mixed alphabet distributions. See Remark 3 following Theorem 1.

The following claims are readily verified.

Lemma 4. *Suppose P_{XY} satisfies property **(P1)**. Then*

(i) $\frac{\partial}{\partial v} F^{-1}_{\Theta|Y}(v \mid y) = 1/f_{\Theta|Y}(F^{-1}_{\Theta|Y}(v \mid y) \mid y)$ *for P_Y-a.a. y.*

(ii) $I(X;Y) = I(\Theta;Y) < \infty$.

Finally, we say that a r.v. X is *stochastically smaller* than another r.v. Y, if $\Pr(Y \le a) \le \Pr(X \le a)$ for any a. More generally, we say that X is stochastically smaller than Y given some event A, if $\Pr(Y \le a \mid A) \le \Pr(X \le a)$ for any a.

B. Setup

We are concerned with the following feedback communication setup. A transmitter is in possession of a *message point* $\Theta_0 \sim \text{Unif}([0,1])$, its binary expansion representing an infinite i.i.d. uniform bit sequence to be reliably communicated to a receiver over a memoryless channel $P_{Y|X}$. The input and output of the channel at time n are denoted X_n and Y_n respectively. We assume there is a noiseless instantaneous feedback link from the receiver back to the transmitter, so that at time n the transmitter is in possession of Y^{n-1}. The memoryless channel model means that Y_n is independent of $(X^{n-1}, Y^{n-1}, \Theta_0)$ given X_n, and that $Y_n \mid X_n = x_n \sim P_{Y|X}(\cdot \mid x_n)$. Furthermore, we assume the transmitter and the receiver share some common randomness; specifically, we assume they can jointly draw an i.i.d. sequence $\{V_n \sim \text{Unif}\,([0,1])\}_{n=1}^{\infty}$, where V_n is statistically independent of $(\Theta_0, X^n, Y^n, V^{n-1})$.

A (sequential, horizon-free) *transmission scheme* is an infinite sequence of mappings that determine the next channel input X_{n+1} as a function of (Θ_0, Y^n, V^n). A *decoding rule* is a corresponding sequence of functions that map (Y^n, V^n) to an interval (modulo 1) J_n, in which the receiver believes the message point lies. The *error probability* attained by a scheme and a decoding rule at time n is $p_e = \Pr(\Theta_0 \notin J_n)$, and the associated *instantaneous rate* is $R_n = -\frac{1}{n}\log|J_n|$. The relation to decoding actual bits is simple: Identifying the said interval of size 2^{-nR_n} essentially guarantees that the nR_n most significant bits of Θ_0 can be decoded with error probability p_e, up to technical edge issues that can be easily resolved (see [1]). A transmission scheme is said to attain a rate R, if for any target error probability $p_e > 0$ there is a suitable decoding rule such that $\Pr(R_n \ge R) \to 1$ as $n \to \infty$. In the following two subsections we describe a simple and optimal construction of a transmission scheme and decoding rule, namely the randomized PM scheme with RIFS decoding.

C. Randomized Posterior Matching

Let $P_{Y|X}$ be a memoryless channel law, and set some input distribution P_X (say, capacity achieving under some input constraint). Consider the following recursively defined transmission scheme:

$$\begin{aligned} \Theta_1 &= \Theta_0 \\ X_n &= F_X^{-1}(\Theta_n) \\ \Theta_{n+1} &= \left(F_{\Theta|Y}\left(\Theta_n \mid Y_n\right) + V_n\right) \mod 1 \end{aligned} \tag{1}$$

The scheme in (1) will be referred to as the *randomized PM scheme*. Note that for $V_n = 0$ this coincides with the classical PM scheme [1]. The randomization idea is key to our simplified analysis, and is due to Li and El Gamal [8] who analyzed a non-sequential fixed-rate fixed-block-length version of this scheme in a DMC setting.

We recall a few known properties of PM that are also inherited by its randomized sibling, with minor modifications accounting for common randomness. The proofs follow easily from the associated claims in [1], e.g. by thinking of (Y_n, V_n) as the channel output, and are omitted.

Lemma 5. *The randomized PM scheme satisfied the following:*

(i) $\Theta_n \sim$ *Unif*$([0,1])$*,* $X_n \sim P_X$*, and* $Y_n \sim P_Y$*.*
(ii) Θ_n *(and hence* X_n*) is statistically independent of* (Y^{n-1}, V^{n-1})*.*
(iii) $\{Y_n\}_{n=1}^{\infty}$ *and* $\{V_n\}_{n=1}^{\infty}$ *are mutually independent i.i.d. sequences.*
(iv) $I(\Theta_0; Y_n \mid Y^{n-1}, V^n) = I(X;Y)$*.*
(v) $I(\Theta_0; Y^n \mid V^n) = nI(X;Y)$*.*

D. Reversed Iterated Function System (RIFS) Decoding

In this subsection we describe a decoding rule for the randomized PM, that maps Y^n into an interval that is guaranteed to contain the message point Θ_0 up to a prescribed error probability (see [1] for more details). Let $F^{-1}_{\Theta|Y}(v \mid y)$ be the inverse PM kernel, i.e.,

$$F^{-1}_{\Theta|Y}(v \mid y) \triangleq \inf\{\theta : F_{\Theta|Y}(\theta \mid y) > v\}.$$

Set some target error probability $p_e > 0$, and let $J_0 \subset (0,1)$ be an interval of size $|J_0| = 1 - p_e$. The RIFS decoder outputs the interval J_n defined recursively by

$$J_{k+1} = F^{-1}_{\Theta|Y}((J_k - V_{n-k}) \bmod 1 \mid Y_{n-k}) \tag{2}$$

for $k = 0, \ldots, n-1$. Recall that we effectively identify $[0,1)$ with the circle $\mathbb{R}/\mathbb{Z}$, hence we allow wrap-around intervals, i.e., the interval (a,b) for $a > b$ is the union $(a,1) \cup [0,b)$.

Lemma 6 ([1])**.** *The probability of error incurred by the above RIFS decoder is* $\Pr(\Theta_0 \notin J_n) = p_e$*.*

Proof:

$$\begin{aligned}
\Pr(\Theta_0 \in J_n) &= \Pr(\Theta_1 \in J_n) \\
&= \mathbb{E}\Pr(\Theta_1 \in J_n \mid Y_1, V_1) \\
&= \mathbb{E}\Pr(\Theta_2 \in J_{n-1} \mid Y_1, V_1) && (3)\\
&= \Pr(\Theta_2 \in J_{n-1}) && (4)\\
&= \cdots && (5)\\
&= \Pr(\Theta_n \in J_0) \\
&= 1 - p_e. && (6)
\end{aligned}$$

(3) follows since (2) is invertible given Y_{n-k}, V_{n-k}, by virtue of property (**P1**). (4) follows since by Lemma 5 Θ_{k+1} is independent of (Y_k, V_k). In (5) we iterate the same arguments, and (6) holds by definition. ■

Define the sequence of *contraction terms*:

$$L_k \triangleq \log\left(\frac{|J_{k-1}|}{|J_k|}\right),$$

and set $L_0 \triangleq -\log(1-p_e)$. Define further

$$R_n \triangleq \frac{1}{n}\sum_{k=0}^{n} L_k.$$

From the discussion above it is clear that the RIFS decoder outputs an interval of (random) size 2^{-nR_n} in which Θ_0 is guaranteed to lie with probability $1 - p_e$. Therefore, R_n is the (random) instantaneous rate of randomized PM under RIFS decoding with error probability p_e. In what follows, we will be interested in guarantees on R_n. As we shall see, in many cases R_n becomes arbitrarily close (for any target p_e) to the optimal value $I(X;Y)$ with high probability as n grows large. Thus, randomized PM can achieve any rate up to channel capacity.

III. Main Result

We state our main result, showing that under very mild regularity conditions the randomized PM scheme with RIFS decoding achieves any rate below the mutual information.

Theorem 1. *Let $(X,Y) \sim P_{XY} \in \mathfrak{F}$ and assume that $0 < I(X;Y) < \infty$. Then for any target error probability p_e and any $\varepsilon > 0$, the decoding rate achieved by the associated randomized PM scheme with RIFS decoding satisfies*

$$\lim_{n\to\infty} \Pr(R_n > I(X;Y) - \varepsilon) = 1$$

Remark 3. The conditions in the theorem are very general, and specifically hold in the following cases:

- For any discrete memoryless channel with any input distribution such that $I(X;Y) > 0$. In this case [1] the PM kernel is a quasi-linear function in θ for any fixed y, with slopes corresponding to the conditional distributions of x given y.
- When the conditional p.d.f. $f_{X|Y}(x|y)$ exists, is bounded, and has bounded support, for any y.
- For any additive noise channel $Y = X + Z$ where Z is independent of X, both Z and Y have bounded p.d.fs, and either:
  - $f_Z(z), f_Y(y)$ have bounded supports; or,
  - $f_Z(z) \geq 2^{-O(|z|^{k_1})}, f_Y(y) \geq 2^{-O(|y|^{k_2})}$ and $\mathbb{E}|Z|^{3k_1}, \mathbb{E}|Y|^{3k_2} < \infty$ for some $k_1, k_2 > 0$. This includes in particular the additive Gaussian channel with a Gaussian input, where the scheme essentially reduces to the well known Schalwijk-Kailath Scheme [11], [12]. Note that this subfamily also includes mixed alphabet channels, e.g. binary input and additive Gaussian noise, etc.

Remark 4. The original PM optimality result (no randomization) requires the posterior matching kernel to be free of any fixed points [1]. It was further shown in [13] that the existence of such fixed points is possible, and that in such a case no positive rate can be attained, unless a suitable input transformation is applied. We note that the randomized PM does not suffer from this issue; the fixed point problem is "washed away" by the random shifting operation.

IV. Proof of Main Result

A. Proof Sketch

Before we proceed to formally prove Theorem 1, we give a heuristic argument that captures the essence of the proof. Let $S_n \triangleq nR_n = \sum_{k=0}^{n} L_k$ be the sum of contraction terms at time n. First, note that if we fix the horizon n, the process $\{S_k\}_{k=1}^{n}$ is a Markov chain in the time index k. Alas, the stochastic process S_n is *not* a Markov chain in the horizon parameter n, since the RIFS process evolves backward in time (see [1] for more details). However, since we are only interested in the asymptotic (marginal) behavior of S_n as the horizon n grows unbounded, then instead of fixing the horizon n and analyzing the process S_k, we can assume the horizon is infinite and think of S_n as a Markov chain for any $n \in \mathbb{N}$ (with some abuse of notations, where we replaced S_k with S_n). The associated processes L_n and J_n will be indexed by n as well. In other words, we are effectively thinking of the decoding process going forward in time, instead of backward.

How does the process S_n evolve? At time n, imagine we are in possession of some random interval J_n of size $|J_n| = 2^{-S_n}$, corresponding to the interval the RIFS holds after n backward iterations. The position of J_n is uniformly distributed over the unit interval modulo 1, due to the random shift operation. We independently draw a r.v. $Y_n \sim P_Y$ (recalling that the output sequence is i.i.d), and apply the inverse PM kernel to obtain the next interval $J_{n+1} = F^{-1}_{\Theta|Y}(J_n \mid Y_n)$, which is then randomly shifted modulo 1. This procedure yields the update

$$S_{n+1} = S_n + L_n, \quad \text{where} \quad L_n = \log\left(\frac{|J_n|}{|J_{n+1}|}\right).$$

The process S_n is thus a Markovian random walk on $\mathbb{R}^+$, starting from $S_0 = -\log(1-p_e)$, with the contraction terms L_n as its increments.

Now, assume that S_n is already very large, i.e. that the associated interval size $|J_n|$ is very small. What is the increment L_n in this case? Clearly, J_n will shrink (or stretch) by a (random) factor that is roughly the derivative

of $F^{-1}_{\Theta|Y}(v \mid y)$ w.r.t. v, evaluated for $y = Y_n$ and at v that is (say) the random midpoint of J_n, which is $\sim \text{Unif}([0,1])$ and independent of Y_n. By Lemma 4 claim (i), this derivative is equal to $1/f_{\Theta|Y}(F^{-1}_{\Theta|Y}(v \mid y))$. The contraction term is hence roughly $\log f_{\Theta|Y}(F^{-1}_{\Theta|Y}(V_n \mid Y_n) \mid Y_n)$. Defining $\tilde{\Theta} = F^{-1}_{\Theta|Y}(V_n \mid Y_n)$, it is readily verified that $(\tilde{\Theta}, Y_n) \sim P_{\Theta Y}$ as induced by P_{XY} and $X = F^{-1}_X(\Theta)$ (see Lemma 7). Thus, we conclude that when S_n is large, the contraction term L_n has distribution close to that of the r.v. $\log f_{\Theta|Y}(\Theta \mid Y)$, and hence $\mathbb{E}L_n \approx I(\Theta;Y) = I(X;Y)$. Thus, as long as S_n does not become too small, it grows like the sum of roughly i.i.d. random variables with expectation $I(X;Y)$, which is why we expect S_n to be close to $nI(X;Y)$.

Of course, the devil is in the details. The main technical challenge is to bound the behavior of the chain for small S_n, in which case the contraction terms behave quite differently; in contrast to the case of a large S_n where the distribution of the contraction terms is essentially independent of the actual value of S_n, here this distribution strongly depends on the exact position of the random walk. More specifically, instead of being the logarithm of the derivative of the inverse PM kernel, the contraction terms in the "small" regime correspond to the logarithm of the λ-smoothed derivative of the inverse PM kernel, with a smoothing factor of $\lambda = 2^{-S_n}$. In the next subsection, we deal with these difficulties: First, we show that S_n spends overall little time in the "small" regime (note that it can go back and forth between "large" and "small"). Then, we couple the process S_n with a simpler process S'_n that has only two modes of i.i.d. behavior, corresponding to whether S_n is "small" or "large". We show that the contribution of the "small" mode of S'_n is negligible, and that consequently S'_n is close to $nI(X;Y)$ with high probability. The proof is then completed by observing that S'_n is stochastically smaller than S_n.

B. Detailed Proof

In this subsection we prove Theorem 1. We use the definition of S_n as a Markovian random walk on $\mathbb{R}^+$, with the time arrow going forward instead of backward, as described in the previous subsection. Define the random variable

$$L^{(\lambda)} \triangleq -\log D_\lambda[F^{-1}_{\Theta|Y}(V \mid Y)], \tag{7}$$

where $Y \sim P_Y$ and $V \sim \text{Unif}([0,1])$ are independent. Clearly, the distribution of $L^{(\lambda)}$ is the same as the distribution of the contraction factor L_n given that $S_{n-1} = -\log\lambda$.

We begin by proving two lemmas characterizing the behavior of $L^{(\lambda)}$.

Lemma 7. *Let $\widetilde{\Theta} \triangleq F^{-1}_{\Theta|Y}(V \mid Y)$. Then $(\widetilde{\Theta}, Y) \sim P_{\Theta Y}$ and*

$$\lim_{\lambda\to 0} L^{(\lambda)} = \log \frac{f_{\Theta|Y}(\widetilde{\Theta} \mid Y)}{f_\Theta(\widetilde{\Theta})} \quad a.s.$$

Proof: By assumption (**P1**), Lemma 2, and Lemma 4 claim (i), we have that given $V = v$ and $Y = y$

$$\begin{aligned}\lim_{\lambda\to 0} -\log D_\lambda[F^{-1}_{\Theta|Y}(v \mid y)] &= -\log \frac{\partial}{\partial v}\left(F^{-1}_{\Theta|Y}(v \mid y)\right) \\ &= \log f_{\Theta|Y}(F^{-1}_{\Theta|Y}(v \mid y) \mid y) \\ &= \log \frac{f_{\Theta|Y}(F^{-1}_{\Theta|Y}(v \mid y) \mid y)}{f_\Theta(F^{-1}_{\Theta|Y}(v \mid y))}\end{aligned}$$

for P_{VY}-a.a. (v,y), where the last step follows trivially since $f_\Theta(\theta) = 1$ for any $\theta \in (0,1)$. It follows that $L^{(\lambda)}$ converges a.s. to the random variable $\log f_{\Theta|Y}(\widetilde{\Theta} \mid Y)$, where $\widetilde{\Theta}$ is defined in the Lemma. Now

$$\begin{aligned}\Pr(\widetilde{\Theta} \le \theta \mid Y = y) &= \Pr(F^{-1}_{\Theta|Y}(V \mid Y) \le \theta \mid Y = y) \\ &= \Pr(V \le F_{\Theta|Y}(\theta \mid y) \mid Y = y) \qquad (8) \\ &= F_{\Theta|Y}(\theta \mid y), \qquad (9)\end{aligned}$$

where (8) holds due to the strict monotonicity of the PM kernel under assumption (**P1**), and (9) follows since Y and V are independent. Hence, $(\widetilde{\Theta}, Y) \sim P_{\Theta Y}$ according to the joint distribution induced by $(P_X, P_{Y|X})$. This completes the proof. ∎

Lemma 8. *$\mathbb{E}L^{(\lambda)}$ satisfies the following properties:*

(i) *$\mathbb{E}L^{(\lambda)}$ is continuous in λ over $[0,1]$.*
(ii) *$\lim_{\lambda\to 1}\mathbb{E}L^{(\lambda)} = 0$.*
(iii) *$\lim_{\lambda\to 0}\mathbb{E}L^{(\lambda)} = I(X;Y)$.*
(iv) *If $I(X;Y) > 0$ then $0 < \mathbb{E}L^{(\lambda)} < I(X;Y)$ for any $\lambda \in (0,1)$.*

Proof: The first claim follows easily from assumption **(P1)**, by the continuity of the inverse PM kernel. The second claim holds since $F^{-1}(\cdot \mid y)$ maps the unit interval to itself for any y. Let us prove the third claim. By property **(P2)** of the family $\mathfrak{F}$, there must exists some $\lambda_0 > 0$ such that $\{L^{(\lambda)}\}_{\lambda\in(0,\lambda_0)}$ is bounded in $\mathcal{L}^p$ for $p = 2+\delta > 1$. Hence $\{L^{(\lambda)}\}_{\lambda\in(0,\lambda_0)}$ are uniformly integrable. By Lemma 7, $L^{(\lambda)}$ also converges a.s. to a finite limit. Thus, by Vitali's convergence theorem [10], we can change the order of limit and expectation, i.e.,

$$\begin{aligned}\lim_{\lambda\to 0}\mathbb{E}L^{(\lambda)} &= \mathbb{E}\lim_{\lambda\to 0} L^{(\lambda)}\\ &= \mathbb{E}\log\frac{f_{\Theta|Y}(\Theta\mid Y)}{f_\Theta(\Theta)}\\ &= I(\Theta;Y)\\ &= I(X;Y),\end{aligned}$$

where we have used Lemma 4 claim (ii) in the last step.

For the fourth claim, note that we can write

$$L^{(\lambda)} = -\log\mathbb{E}_Q\left(1/f_{\Theta|Y}(F^{-1}_{\Theta|Y}((V+Q)\bmod 1\mid Y)\mid Y)\right),$$

where $Q\sim\text{Unif}([-\frac{\lambda}{2},\frac{\lambda}{2}])$ is independent of V,Y. We therefore have that

$$\begin{aligned}\mathbb{E}L^{(\lambda)} &= \mathbb{E}_{V,Y}L^{(\lambda)}\\ &= -\mathbb{E}_{V,Y}\log\mathbb{E}_Q\left(1/f_{\Theta|Y}(F^{-1}_{\Theta|Y}((V+Q)\bmod 1\mid Y)\mid Y)\right)\\ &< \mathbb{E}_{V,Y}\mathbb{E}_Q\log f_{\Theta|Y}(F^{-1}_{\Theta|Y}((V+Q)\bmod 1\mid Y)\mid Y) \qquad (10)\\ &= \mathbb{E}_{V',Y}\log f_{\Theta|Y}(F^{-1}_{\Theta|Y}(V'\mid Y)\mid Y)\\ &= \mathbb{E}_{\Theta Y}\log f_{\Theta|Y}(\Theta\mid Y) \qquad (11)\\ &= I(\Theta;Y)\\ &= I(X;Y), \qquad (12)\end{aligned}$$

where $V' = (V+Q)\bmod 1$ is uniform over the unit interval. We have used Jensen's inequality in (10), which is strict since $\lambda > 0$ and $I(\Theta;Y) > 0$. (11) follows from Lemma 7, and (12) follows again from Lemma 4 claim (ii). Similarly,

$$\begin{aligned}\mathbb{E}L^{(\lambda)} &= -\mathbb{E}_{V,Y}\log\mathbb{E}_Q\left(1/f_{\Theta|Y}(F^{-1}_{\Theta|Y}((V+Q)\bmod 1\mid Y)\mid Y)\right)\\ &> -\log\mathbb{E}_{V,Y}\mathbb{E}_Q\left(1/f_{\Theta|Y}(F^{-1}_{\Theta|Y}((V+Q)\bmod 1\mid Y)\mid Y)\right) \qquad (13)\\ &= -\log\mathbb{E}_{V',Y}\left(1/f_{\Theta|Y}(F^{-1}_{\Theta|Y}(V'\mid Y)\mid Y)\right)\\ &= -\log\mathbb{E}_{\Theta Y}\left(1/f_{\Theta|Y}(\Theta\mid Y)\right)\\ &= -\log\mathbb{E}_Y\mathbb{E}_{\Theta|Y}\left(1/f_{\Theta|Y}(\Theta\mid Y)\right)\\ &= -\log\mathbb{E}_Y 1\\ &= 0. \qquad (14)\end{aligned}$$

■

Using the properties of $L^{(\lambda)}$ established above, we would like to show that S_n spends little time close to the origin. To that end, we first prove a the following lemma.

Lemma 9. *S_n is a submatrigale on $\mathbb{R}^+$, and $\Pr(\limsup_{n\to\infty} S_n = \infty) = 1$.*

Proof: The submartingale claim follows immediately from Lemma 8 property (iv). Let us prove the other claim. Recall that by Lemma 8, $\mathbb{E}L^{(\lambda)}$ is a continuous function of λ over $[0,1]$, and $0 < \mathbb{E}L^{(\lambda)} < I(X;Y)$ for any $\lambda \in (0,1]$, where the upper and lower bounds are approached as λ tends to zero and one respectively. It is therefore easy to construct a two-sided monotonically decreasing sequence $\{\lambda_k\}_{k=-\infty}^{\infty}$ with $\lim_{k\to-\infty}\lambda_k = 1$ and $\lim_{k\to\infty}\lambda_k = 0$ such that

$$\inf_{\lambda\in[\lambda_{k+1},\lambda_k)} \mathbb{E}L^{(\lambda)} > 3\log\frac{\lambda_k}{\lambda_{k+1}} \tag{15}$$

for any k. Hence,

$$\begin{aligned}
\delta_k &\triangleq \inf_{\lambda\in[\lambda_{k+1},\lambda_k)} \Pr\left(L^{(\lambda)} > 2\log\frac{\lambda_k}{\lambda_{k+1}}\right) \\
&\geq \inf_{\lambda\in[\lambda_{k+1},\lambda_k)} \Pr\left(L^{(\lambda)} > \frac{2}{3}\inf_{\lambda'\in[\lambda_{k+1},\lambda_k)} \mathbb{E}L^{(\lambda')}\right) &(16)\\
&\geq \inf_{\lambda\in[\lambda_{k+1},\lambda_k)} \Pr\left(L^{(\lambda)} > \frac{2}{3}\mathbb{E}L^{(\lambda)}\right) &(17)\\
&> 0, &(18)
\end{aligned}$$

where (16) follows from (15), choosing $\lambda' = \lambda$ establishes (17), and (18) trivially holds since $\mathbb{E}L^{(\lambda)} > 0$ on any closed subinterval of $(0,1)$.

Let $\{\tau_{j,k}\}_{j=1}^{T_k}$ be the sequence of all time indices n where $S_n \in (-\log\lambda_k, -\log\lambda_{k+1}]$, where T_k is the (possibly infinite) total number of such occurrences. Let M_k be the maximal time index n for which $S_n > -\log\lambda_{k+1}$, and let b be some fixed positive integer.

$$\begin{aligned}
\Pr\left(\limsup_{n\to\infty} S_n \in (-\log\lambda_k, -\log\lambda_{k+1}]\right) &= \Pr\left(M_k < \infty, T_k = \infty\right) \\
&\leq \Pr\left(M_k < \infty, T_k \geq M_k + b\right) \\
&= \sum_{m=0}^{\infty} \Pr\left(T_k \geq m + b \mid M_k = m\right)\Pr(M_k = m) \\
&\leq \sum_{m=0}^{\infty} \Pr\left(L_{\tau_{j,k}} \leq \log\frac{\lambda_k}{\lambda_{k+1}}, m < j \leq m + b \mid M_k = m\right)\Pr(M_k = m) \\
&\leq \sum_{m=0}^{\infty} (1-\delta_k)^b \Pr(M_k = m) \\
&\leq (1-\delta_k)^b.
\end{aligned}$$

Since $\delta_k > 0$, and as the above upper bound holds for any b and k, it must be that

$$\Pr\left(\limsup_{n\to\infty} S_n \in (-\log\lambda_k, -\log\lambda_{k+1}]\right) = 0.$$

The proof is now concluded by noting that $\mathbb{R}^+ = \bigcup_k (-\log\lambda_k, -\log\lambda_{k+1}]$. ∎

We now further strengthen Lemma 9 and show that S_n in fact diverges a.s., which will specifically show that it spends little time below any threshold t. Let $N_{t,n}$ be the number of times S_k falls below t until time n, i.e.,

$$N_{t,n} \triangleq \sum_{k=1}^{n} \mathbb{1}(S_k < t),$$

and let $N_t \triangleq \lim_{n\to\infty} N_{t,n}$ be a random variable on $\mathbb{N}\cup\{\infty\}$.

Lemma 10. *$S_n \to \infty$ almost surely, hence $\Pr(N_{t,n} > m) \leq \Pr(N_t > m) = \delta(m)$ where $\delta(m) \to 0$ as $m \to \infty$.*

Proof: The proof is based on arguments similar to [14]. Consider the process $T_n = 1 - \frac{1}{1+S_n}$. Below we show that T_n converges a.s., which together with Lemma 9 implies that that $T_n \to 1$ a.s. and hence $S_n \to \infty$ a.s., establishing the lemma.

First, we show it is sufficent to prove that there exists some $t_0 \in (0,1)$ such that $\mathbb{E}(T_{n+1} \mid T_n = t) \geq t$ for any $t \geq t_0$. To see that, define the process $T'_n = \max(T_n, t_0)$, and note that by definition it holds that $\mathbb{E}(T'_{n+1} \mid T'_n = t) \geq t$ for any t, hence T'_n is a submartingale. Moreover, $\mathbb{E}|T'_n| \leq 1$ for all n. By Lemma 1, it must therefore be that T'_n convergences a.s. to a limit. Since $\Pr(\limsup_{n\to\infty} T'_n = 1) \geq \Pr(\limsup_{n\to\infty} T_n = 1) = 1$, this limit must be 1, i.e., $T'_n \to 1$ a.s . Since $T_n = T'_n$ whenever $T'_n \geq t_0$, it must be that $T_n \to 1$ a.s. as well.

It remains to show the existence of such a t_0. Let us first establish some guarantees on the first and second moments of $L^{(\lambda)}$, conditioned on an event that $L^{(\lambda)} > a$ for some a. From Lemma 8 we know that $\mathbb{E}L^{(\lambda)}$ approaches $I(X;Y) > 0$ continuously as $\lambda \to 0$, hence in particular there is some $c_1 > 0$ such that $\mathbb{E}L^{(\lambda)} > c_1$ for all $\lambda > 0$ small enough. Trivially, it also holds that for any a

$$\mathbb{E}\left(L^{(\lambda)} \mid L^{(\lambda)} > a\right) \geq \mathbb{E}L^{(\lambda)} > c_1 > 0 \tag{19}$$

for any $\lambda > 0$ small enough. Moreover, property (**P2**) of the family $\mathfrak{F}$ implies that $L^{(\lambda)}$ is uniformly bounded in $\mathcal{L}^2$ for all $\lambda > 0$ small enough, hence $\mathbb{E}|L^{(\lambda)}|^2 < c_2$ for some some $c_2 < \infty$. Trivially then, for any a it also holds that

$$\Pr(L^{(\lambda)} > a) \cdot \mathbb{E}\left(|L^{(\lambda)}|^2 \mid L^{(\lambda)} > a\right) \leq \mathbb{E}\left(|L^{(\lambda)}|^2\right) < c_2 < \infty \tag{20}$$

for all $\lambda > 0$ small enough.

Now, define the function $g(s,\ell) \triangleq \frac{1}{1+s} - \frac{1}{1+s+\ell}$. Since the process S_n is nonnegative, we can clearly limit our discussion to $\ell \geq -s$, and hence to $g(s,\ell) \geq -1$. Let us write

$$\begin{aligned} g(s,\ell) &= \frac{\ell}{(1+s)^2 + \ell(1+s)} \\ &= \frac{\ell}{(1+s)^2} - \frac{\ell^2}{(1+s)^3 + \ell(1+s)^2}. \end{aligned}$$

Setting any $\alpha \in (0,1)$, it therefore holds that for any $\ell \geq -(1+s)^\alpha$ and $s > 2^{\frac{1}{1-\alpha}} - 1$,

$$\begin{aligned} g(s,\ell) &\geq \frac{\ell}{(1+s)^2} - \frac{\ell^2}{(1+s)^3 - (1+s)^{2+\alpha}} \\ &\geq \frac{\ell}{(1+s)^2} - \frac{\ell^2}{2(1+s)^3}. \end{aligned} \tag{21}$$

Our analysis will now naturally depend on the event $L_n \geq -(1+s)^\alpha$. Let us first upper bound the probability of the complementary event:

$$\begin{aligned} \Pr(L_n < -(1+s)^\alpha \mid S_n = s) &\leq \Pr(|L_n| > (1+s)^\alpha \mid S_n = s) \\ &= \Pr(|L_n|^{2+\delta} > (1+s)^{\alpha(2+\delta)} \mid S_n = s) \\ &\leq \frac{\mathbb{E}\left(|L_n|^{2+\delta} \mid S_n = s\right)}{(1+s)^{\alpha(2+\delta)}} \qquad (22) \\ &= \frac{\mathbb{E}\left(\left|L^{(2^{-s})}\right|^{2+\delta}\right)}{(1+s)^{\alpha(2+\delta)}} \\ &\leq c_3 \cdot (1+s)^{-\alpha(2+\delta)} \qquad (23) \end{aligned}$$

for some $c_3 > 0$ and any s large enough. We used Markov's inequality in (22), and (23) is again by virtue of property (**P2**) of the family $\mathfrak{F}$, that implies $L^{(\lambda)}$ is uniformly bounded in $\mathcal{L}^{2+\delta}$ for all $\lambda > 0$ small enough.

Writing $t = 1 - \frac{1}{1+s}$ we have that for any s sufficiently larger than $2^{\frac{1}{1-\alpha}} - 1$

$$\begin{aligned}
\mathbb{E}(T_{n+1} - T_n \mid T_n = t) &= \mathbb{E}\left(g(s, L_n) \mid S_n = s\right) \\
&= \mathbb{E}\left(g(s, L^{(2^{-s})})\right) \\
&= \Pr\left(L^{(2^{-s})} < -(1+s)^{\alpha}\right) \cdot \mathbb{E}\left(g(s, L^{(2^{-s})}) \mid L^{(2^{-s})} < -(1+s)^{\alpha}\right) \\
&\quad + \Pr\left(L^{(2^{-s})} \geq -(1+s)^{\alpha}\right) \cdot \mathbb{E}\left(g(s, L^{(2^{-s})}) \mid L^{(2^{-s})} \geq -(1+s)^{\alpha}\right) \\
&\geq -c_3 \cdot (1+s)^{-\alpha(2+\delta)} \qquad (24) \\
&\quad + \Pr\left(L^{(2^{-s})} \geq -(1+s)^{\alpha}\right) \cdot \mathbb{E}\left(\frac{L^{(2^{-s})}}{(1+s)^2} \mid L^{(2^{-s})} \geq -(1+s)^{\alpha}\right) \\
&\quad - \Pr\left(L^{(2^{-s})} \geq -(1+s)^{\alpha}\right) \cdot \mathbb{E}\left(\frac{\left|L^{(2^{-s})}\right|^2}{2(1+s)^3} \mid L^{(2^{-s})} \geq -(1+s)^{\alpha}\right) \\
&\geq -c_3 \cdot (1+s)^{-\alpha(2+\delta)} + \left(1 - c_3 \cdot (1+s)^{-\alpha(2+\delta)}\right) \cdot \frac{c_1}{(1+s)^2} - \frac{c_2}{2(1+s)^3}. \qquad (25)
\end{aligned}$$

(24) follows from (21), (23), and since $g(s, \ell) \geq -1$. (25) follows from (19), (20), and (23). Examining (25) for any $\frac{2}{2+\delta} < \alpha < 1$, it is immediately clear that this lower bound on the expected increment is positive for all large enough s, and hence for all t sufficiently close to 1. This concludes the proof.

■

After establishing that $S_n \to \infty$ a.s., we would like to further determine how fast this happens. To that end, we will define a coupled process S'_n that will be easier to handle, and will be stochastically smaller than S_n. Loosely speaking, S'_n will have two modes of i.i.d. random walk behavior corresponding to whether S_n is above or below the threshold t; it will also grow slower than S_n in each of these regimes.

To do that, we first define two random variables U, W that will be stochastically smaller than L_n given that S_n is above or below the threshold t respectively, and will later determine the increments of the coupled process S'_n in these two regimes. For brevity, we omit the dependence of U, W on t. Recall the definition of $L^{(\lambda)}$ in (7). We first define $\widetilde{U}, \widetilde{W}$ via their c.d.fs as follows:

$$\Pr(\widetilde{U} \leq u) \triangleq \sup_{\lambda \in (0, 2^{-t}]} \Pr\left(L^{(\lambda)} \leq u\right),$$

$$\Pr(\widetilde{W} \leq w) \triangleq \sup_{\lambda \in (2^{-t}, 1)} \Pr\left(L^{(\lambda)} \leq w\right).$$

Now, setting some large number $\xi > 0$, we define U, W as the truncation of $\tilde{U}, \tilde{W}$:

$$U \triangleq \min(\widetilde{U}, \xi), \quad W \triangleq \min(\widetilde{W}, \xi).$$

Again, the dependence on ξ will be omitted for notational clarity. The following lemma describes some important properties of U and W. The proof is relegated to the appendix.

Lemma 11. *The following properties hold:*

(i) U *is stochastically smaller than* L_n *given* $S_{n-1} = t_0$ *for any* $t_0 \geq t$
(ii) W *is stochastically smaller than* L_n *given* $S_{n-1} = t_0$ *for any* $t_0 < t$
(iii) $\mathbb{E}U \leq I(X;Y)$ *for any* t, ξ.
(iv) $\lim_{\xi \to \infty} \lim_{t \to \infty} \mathbb{E}U = I(X;Y)$.
(v) $\mathbb{E}|W| < \infty$ *for any* $\xi, t > 0$.

We are now ready to define the coupled process S'_n. Let $\{U_n\}$ and $\{W_n\}$ be two i.i.d. sequences with distributions P_U and P_W respectively, such that the processes $\{U_n\}, \{W_n\}, \{S_n\}$ are mutually independent.

Define S'_n to be the random walk process generated by replacing the increments of the process S_n process with U or W elements, according to whether S_n is above or below the threshold. Precisely:

$$S'_n = \sum_{k=1}^{n-N_{t,n}} U_k + \sum_{k=1}^{N_{t,n}} W_k.$$

Note that unlike S_n, the coupled process S'_n can become negative, since $\Pr(W \leq 0) = 1$. Also, S'_n does not contain the fixed initialization term $L_0 = -\log(1-p_e)$. The proof of the following lemma appears in the appendix.

Lemma 12. *S'_n is stochastically smaller than S_n.*

Let us now show the probability S'_n falls below $n(I(X;Y)-\varepsilon)$ vanishes with n.

Lemma 13. $\lim_{n\to\infty} \Pr(S'_n > n(I(X;Y)-\varepsilon)) = 1$ *for any* $\varepsilon > 0$.

Proof: We write $I = I(X;Y)$ for short. Set ξ and t large enough so that such that

$$I - \mathbb{E}U \leq \varepsilon/8, \tag{26}$$

which is possible by virtue of Lemma 11 claims (iv) and (iii). Then:

$$\begin{aligned}
\Pr(S'_n < n(I-\varepsilon)) &= \Pr\left(\sum_{k=1}^{n-N_{t,n}} U_k + \sum_{k=1}^{N_{t,n}} W_k < n(I-\varepsilon)\right) \\
&\leq \Pr(N_{t,n} > m) + \sum_{r=1}^{m} \Pr(N_{t,n}=r) \Pr\left(\sum_{k=1}^{n-r} U_k + \sum_{k=1}^{r} W_k < n(I-\varepsilon) \mid N_{t,n}=r\right) \\
&\leq \delta(m) + \sum_{r=1}^{m} \Pr(N_{t,n}=r) \left[\Pr\left(\sum_{k=1}^{n-r} U_k < nI - \frac{n\varepsilon}{2} \bigvee \sum_{k=1}^{r} W_k < -\frac{n\varepsilon}{2}\right)\right] \qquad (27) \\
&\leq \delta(m) + \sum_{r=1}^{m} \Pr(N_{t,n}=r) \left[\Pr\left(\sum_{k=1}^{n-r} U_k < nI - \frac{n\varepsilon}{2}\right) + \Pr\left(\sum_{k=1}^{r} W_k < -\frac{n\varepsilon}{2}\right)\right] \qquad (28) \\
&\leq \delta(m) + \sum_{r=1}^{m} \Pr(N_{t,n}=r) \left[\Pr\left(\frac{1}{n-r}\sum_{k=1}^{n-r} U_k < \frac{I-\frac{\varepsilon}{2}}{1-\frac{r}{n}}\right) + \Pr\left(\frac{1}{r}\sum_{k=1}^{r} W_k < -\frac{n\varepsilon}{2r}\right)\right]. \qquad (29)
\end{aligned}$$

(27) follows from Lemma 10 and since the sequences $\{U_n\}, \{V_n\}$ are mutually independent of $\{S_n\}$, hence of $N_{t,n}$ as well. (28) follows from the union bound. Analyzing the first term inside the parenthesis in (29), we note that for any $1 \leq r \leq m$ and $n > m$ large enough,

$$\begin{aligned}
\Pr\left(\frac{1}{n-r}\sum_{k=1}^{n-r} U_k < \frac{I-\varepsilon/2}{1-\frac{r}{n}}\right) &\leq \Pr\left(\frac{1}{n-r}\sum_{k=1}^{n-r} U_k < I - \varepsilon/4\right) \\
&\leq \Pr\left(\frac{1}{n-r}\sum_{k=1}^{n-r} U_k < \mathbb{E}U - \varepsilon/8\right) \qquad (30) \\
&= o_{m,t,\xi,\varepsilon}(1), \qquad (31)
\end{aligned}$$

where (30) follows from (26), and (31) is by virtue of the law of large numbers. Furthermore,

$$\begin{aligned}
\Pr\left(\frac{1}{r}\sum_{k=1}^{r} W_k < -\frac{n\varepsilon}{2r}\right) &\leq \Pr\left(\frac{1}{r}\sum_{k=1}^{r} |W_k| > \frac{n\varepsilon}{2m}\right) \\
&\leq \frac{2m}{n\varepsilon} \cdot \mathbb{E}|W| \qquad (32) \\
&= O_{m,t,\xi,\varepsilon}(n^{-1}), \qquad (33)
\end{aligned}$$

where (32) follows from Markov's inequality, and (33) is by virtue of Lemma 11 property (v). We therefore obtain that for any m and ε there are t, ξ large enough such that

$$\Pr(S'_n < n(I-\varepsilon)) \leq \delta(m) + o_{m,t,\xi,\varepsilon}(1),$$

where $\delta(m) \to 0$ as $m \to \infty$. Since we can fix m arbitrarily large we have that

$$\lim_{n\to\infty} \Pr(S'_n < n(I-\varepsilon)) = 0$$

as desired. ∎

Finally, combining Lemmas 12 and 13 with the definition of R_n, we obtain

$$\begin{aligned}\lim_{n\to\infty} \Pr(R_n > I(X;Y) - \varepsilon) &= \lim_{n\to\infty} \Pr(S_n > n(I(X;Y)-\varepsilon)) \\ &\geq \lim_{n\to\infty} \Pr(S'_n > n(I(X;Y)-\varepsilon)) \\ &= 1,\end{aligned}$$

establishing the theorem.

APPENDIX

Proof of Lemma 3: Define the function $\phi : \mathbb{R} \to \mathbb{R}$

$$\phi(x) = g'(t \bmod 1) \cdot \mathbb{1}(x \in [-1, 2]).$$

Let $M\phi(x)$ be the Hardy-Littlewood maximal function [10, Chapter 7] pertaining to $\phi(x)$, i.e.,

$$\begin{aligned}M\phi(x) &\triangleq \sup_{\lambda>0} \frac{1}{\lambda} \int_{x-\lambda/2}^{x+\lambda/2} |\phi(t)|dt \\ &= \sup_{\lambda>0} \mathbb{E}|\phi(x+Q_\lambda)|, \end{aligned} \tag{34}$$

where $Q_\lambda \sim \text{Unif}\left(\left[-\frac{\lambda}{2}, \frac{\lambda}{2}\right]\right)$. For any $x \in [0,1)$ we can also write

$$\begin{aligned}M\phi(x) &\geq \sup_{\lambda\in(0,1)} \mathbb{E}|\phi(x+Q_\lambda)| \\ &= \sup_{\lambda\in(0,1)} \mathbb{E}|g'((x+Q_\lambda) \bmod 1)| \\ &= \overline{D}[g(x)], \end{aligned} \tag{35}$$

where we have used Lemma 2 in (35). Hence

$$\Pr(\overline{D}[g(x)] > a) \leq \Pr(M\phi(X) > a). \tag{36}$$

The Hardy-Littlewood maximal inequality [10, Chapter 7] implies that for any $a > 0$, the following measure-theoretic "generalized Markov inequality" holds:

$$\begin{aligned}|\{x : M\phi(x) > a\}| &\leq 3a^{-1} \int_{-\infty}^{\infty} |\phi(x)|dx \\ &= 3a^{-1} \int_{-1}^{3} |\phi(x)|dx \\ &= 9a^{-1} \int_0^1 |g'(x)|dx.\end{aligned}$$

Thus, if $X \sim \text{Unif}([0,1])$ then

$$\Pr(M\phi(x) > a) \leq 9a^{-1}\mathbb{E}\left|g'(X)\right|. \tag{37}$$

The proof now follows from (36) and (37). ∎

Proof of Lemma 11:

(i)

$$\begin{aligned}\Pr(L_n \leq u \mid S_{n-1} = t_0) &= \Pr(L^{(2^{-t_0})} \leq u) \\ &\leq \sup_{\lambda\in(0,2^{-t}]} \Pr(L^{(\lambda)} \leq u) \\ &= \Pr(\widetilde{U} \leq u) \\ &\leq \Pr(U \leq u).\end{aligned}$$

(ii) Follows similarly.
(iii) Follows similarly to Lemma 8 claim (iv).
(iv) Follows similarly to Lemma 8 claim (iii).
(v) Write $q(v,y) \triangleq \frac{\partial}{\partial v}\left(F^{-1}_{\Theta|Y}(v \mid y)\right)$, and note that

$$\begin{aligned}\mathbb{E}_Y\mathbb{E}_V|q(V,Y)| &= \mathbb{E}_Y\mathbb{E}_V q(V,Y)\\ &= \mathbb{E}_Y\left(F^{-1}_{\Theta|Y}(1 \mid Y) - F^{-1}_{\Theta|Y}(0 \mid Y)\right)\\ &= 1. \end{aligned} \tag{38}$$

Now, let $w > 0$.

$$\begin{aligned}\Pr(W \le -w) &= \sup_{\lambda\in(2^{-t},1)} \Pr\left(L^{(\lambda)} \le -w\right)\\ &\le \sup_{\lambda\in(2^{-t},1)} \Pr\left(\inf_{\lambda'\in(0,1)} L^{(\lambda')} \le -w\right)\\ &= \Pr\left(\log \sup_{\lambda'\in(0,1)} D_{\lambda'}[F^{-1}_{\Theta|Y}(V \mid Y)] > w\right)\\ &= \Pr\left(\log \overline{D}[F^{-1}_{\Theta|Y}(V \mid Y)] > w\right)\\ &= \Pr\left(\overline{D}[F^{-1}_{\Theta|Y}(V \mid Y)] > 2^w\right)\\ &= \mathbb{E}_Y\left(\Pr\left(\overline{D}[F^{-1}_{\Theta|Y}(V \mid Y)] > 2^w \mid Y\right)\right)\\ &\le 9\cdot 2^{-w}\cdot \mathbb{E}_Y\mathbb{E}_V\,|g(V,Y)| \qquad (39)\\ &= 9\cdot 2^{-w}, \qquad (40)\end{aligned}$$

where in (39) we have used Lemma 3 together with property (**P1**), and (40) follows from (38). Now,

$$\begin{aligned}\mathbb{E}|W| &= \mathbb{E}\left(\int_0^\infty \mathbb{1}(W \ge w)dw + \int_0^\infty \mathbb{1}(W \le -w)dw\right)\\ &= \int_0^\infty \Pr(W \ge w)dw + \int_0^\infty \Pr(W \le -w)dw\\ &\le \int_0^\infty \mathbb{1}(w \le \xi)dw + \int_0^\infty 9\cdot 2^{-w}dw\\ &= \xi + 9\log e.\end{aligned}$$

Note that the bound is independent of t.

■

Proof of Lemma 12: Let $A_n \triangleq \mathbb{1}(S_n < t) = \mathbb{1}\left(\sum_{k=0}^n L_k < t\right)$. For any μ:

$$\begin{aligned}\Pr(S_n < \mu) &\le \Pr\left(\sum_{k=1}^n L_k < \mu\right)\\ &= \mathbb{E}_{L^{n-1}}\Pr\left(\sum_{k=1}^n L_k < \mu \mid L^{n-1}\right)\\ &= \mathbb{E}_{L^{n-1}}\Pr\left(L_n < \mu - \sum_{k=1}^{n-1} L_k \mid L^{n-1}\right)\\ &\le \mathbb{E}_{L^{n-1}}\Pr\left((1-A_{n-1})U_1 + A_{n-1}W_1 < \mu - \sum_{k=1}^{n-1} L_k \mid L^{n-1}\right) \qquad (41)\\ &= \Pr\left((1-A_{n-1})U_1 + A_{n-1}W_1 + \sum_{k=1}^{n-1} L_k < \mu\right),\end{aligned}$$

where (41) follows since (U_1, W_1) are independent of L^{n-1}, and by virtue of the stochastic lower bound properties (i) and (i) in Lemma 11, according to whether U_1 or W_1 is selected by A_{n-1}. Iterating the same argument we obtain

$$\begin{aligned}\Pr(S_n < \mu) &\leq \Pr\left(\sum_{k=1}^{n}(1-A_{n-k})U_k + \sum_{k=1}^{n} A_{n-k}W_k < \mu\right) \\ &= \Pr(S'_n < \mu),\end{aligned}$$

where the last equality follows by noting that $A_k = N_{t,k} - N_{t,k-1}$. This concludes the proof of the Lemma. ■

REFERENCES

[1] O. Shayevitz and M. Feder, "Optimal feedback communication via posterior matching," *IEEE Trans. Info. Theory*, vol. 57, no. 3, pp. 1186 –1222, March 2011.

[2] O. Shayevitz and M. Feder, "Communication with feedback via posterior matching," in *Proc. of the International Symposium on Information Theory*, June 2007.

[3] O. Shayevitz and M. Feder, "The posterior matching feedback scheme: Capacity achieving and error analysis," in *Proc. of the International Symposium on Information Theory*, July 2008.

[4] M. Horstein, "Sequential transmission of digital information with feedback," *Technical report 375, Research Laboratory of Electronics, MIT*, 1960.

[5] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Info. Theory*, vol. IT-9, pp. 136–143, Jul 1963.

[6] T. P. Coleman, "A stochastic control viewpoint on posterior matching-style communication schemes," in *Proc. of the International Symp. on Info. Theory*, Jun 2009.

[7] M. Naghshvar, T. Javidi, and M. Wigger, "Extrinsic Jensen-Shannon divergence: Applications to variable-length coding," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 2148–2164, April 2015.

[8] C. T. Li and A. El Gamal, "An efficient feedback coding scheme with low error probability for discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 2953–2963, June 2015.

[9] J. L. Doob, *Stochastic Processes*, Wiley Publications in Statistics. John Wiley & Sons, 1953.

[10] W. Rudin, *Real and complex analysis*, Mathematics series. McGraw-Hill, 1987.

[11] J. P. M. Schalkwijk and T. Kailath, "A coding scheme for additive noise channels with feedback part I: No bandwidth constraint," *IEEE Trans. Info. Theory*, vol. IT-12, pp. 172 – 182, Apr. 1966.

[12] J. P. M. Schalkwijk, "A coding scheme for additive noise channels with feedback part II: Band-limited signals," *IEEE Trans. Info. Theory*, vol. IT-12, pp. 183 – 189, Apr. 1966.

[13] O. Shayevitz, "Posterior matching variants and fixed-point elimination," in *47th Annual Allerton Conference on Communication, Control, and Computing, 2009.*, 2009, pp. 935–939.

[14] J. Lamperti, "Criteria for the recurrence or transience of stochastic process. I," *Journal of Mathematical Analysis and applications*, vol. 1, no. 3, pp. 314–330, 1960.